

Energy Landscape for Fold-Switching in Regulatory Protein RfaH

Jerelle A. Joseph,[†] Debayan Chakraborty,[‡] and David J. Wales^{*,†}

[†]*Department of Chemistry, University of Cambridge, Lensfield Road,
Cambridge CB2 1EW, United Kingdom*

[‡]*Department of Chemistry, University of Texas at Austin, Austin,
Texas 78712, USA*

E-mail: dw34@cam.ac.uk

Abstract

The C-terminal domain (CTD) of bacterial regulatory protein RfaH undergoes a dramatic structural rearrangement from an α -helical hairpin to a β -barrel. We employ a quasi-continuous interpolation scheme and geometry optimisation techniques to construct a kinetic transition network for this process. Our computed free energy landscape at 310 K is multifunnelled, and the predicted free energy ensembles are in good agreement with experiment and other simulation studies. We find that rearrangement from the α -helical conformer to the β -sheet proceeds via an essentially unstructured state. The techniques refined for the present system should be transferable to other protein conformational switches, with the potential to advance our understanding of such systems.

1 Introduction

Over the last two decades there has been increasing evidence of fold-switching, where certain proteins are capable of adopting distinct, stable folds in a reversible fashion.¹⁻⁸ These proteins, commonly referred to as metamorphic,⁵ extend the classical view of protein conformational dynamics, beyond movements of loop regions and side-chains, to large-scale rearrangements at the level of secondary structure. For example, human chemokine lymphotactin (Ltn) exists as two distinct conformations: a monomeric form (Ltn10), consisting of a three-stranded β -sheet and an α -helix, and an all- β -sheet dimeric form (Ltn40).⁶ Under physiological conditions, the two conformers are in equilibrium, and bind to different molecular partners. Other well-known examples of metamorphic proteins include mitotic arrest deficiency 2 (Mad2) protein² and chloride intracellular channel 1 (CLIC1) protein.¹

Perhaps the most dramatic example of protein conformational switching has been reported for RfaH (162 amino acids; named after the *rfaH* gene that encodes it).⁹ RfaH is a regulatory protein found in *Escherichia coli* (*E.coli*) and *Salmonella*,¹⁰ and is known to increase the expression of genes in operons containing an operon polarity suppressor (*ops*) site (a short, well-conserved DNA sequence).¹⁰⁻¹² RfaH comprises two domains connected by a flexible linker: an N-terminal domain (NTD) and a C-terminal domain (CTD).¹³ In the domain-closed state, the CTD adopts an α -helical hairpin fold, and binds tightly to the NTD.^{13,14} When the transcribing RNA polymerase (RNAP) pauses at the *ops* site, interactions between RNAP, the *ops* site and RfaH lead to domain separation.¹⁴ At this stage, RfaH-NTD binds to RNAP in a clamp-like fashion, modifying RNAP into a pause-resistant state, and ensuring that synthesis of messenger RNA (mRNA) is complete without pausing or premature termination. Accordingly, the main purpose of the CTD in the domain-closed state is to mask the RNAP binding site of RfaH-NTD (a hydrophobic cavity);¹⁴ thus, RfaH-CTD serves as a regulator of transcription, and effectively restricts RfaH to operons containing an *ops* site.

Upon domain separation, the CTD of RfaH undergoes a dramatic conformational tran-

sition: the α -helical hairpin refolds into a five-stranded β -barrel scaffold (i.e. an all- $\alpha \rightarrow$ all- β transition).¹⁴ RfaH-CTD, in the β -barrel conformation, then binds to ribosomal protein S10, thereby recruiting the ribosomal 30S subunit to the nascent mRNA, significantly promoting translation.¹⁴ Hence, for RfaH-CTD the same amino acid sequence folds into two distinct conformations with two distinct functions, constituting a special type of metamorphic system known as a transformer protein¹⁵ (Figure 1).

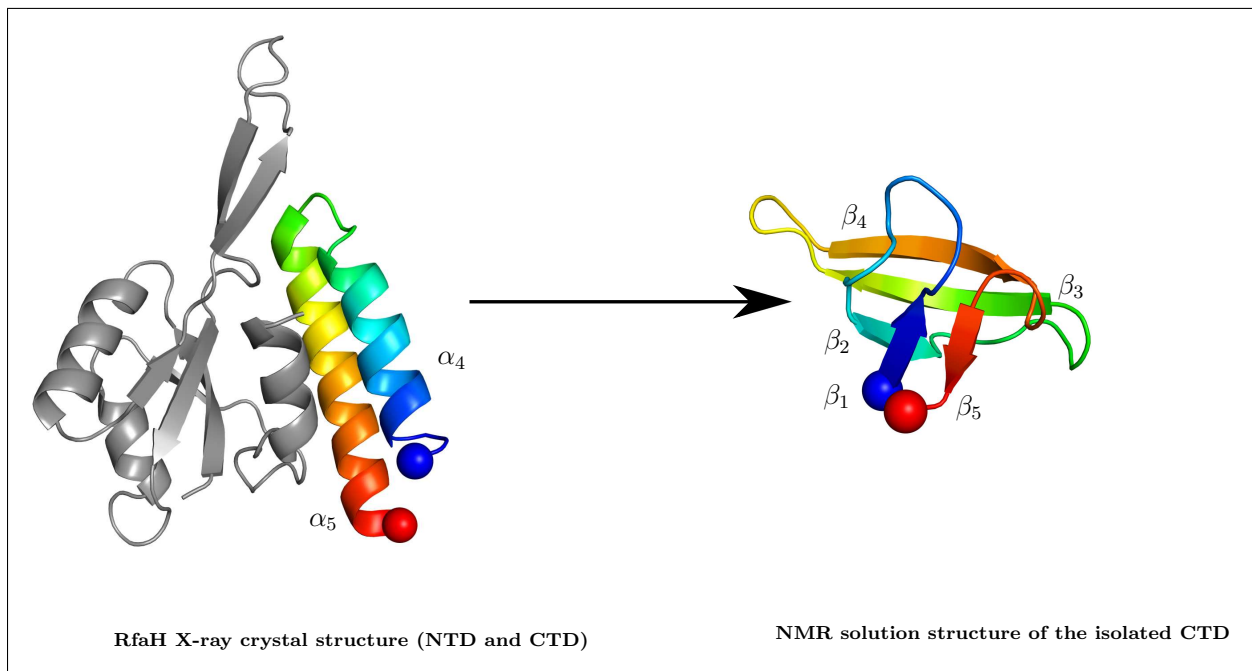


Figure 1: X-ray crystallography structure of RfaH (residues 1–100 and 115–156), and NMR solution structure of the C-terminal domain of RfaH (residues 115–162). Upon domain separation, the CTD of RfaH transforms: from an α -helical hairpin ($\alpha_4 = 115$ –130; $\alpha_5 = 135$ –155) to a five-stranded β -barrel scaffold: β_5 (158–160), β_1 (115–118), β_2 (127–130), β_3 (138–144) and β_4 (149–155). The N-terminus and C-terminus are highlighted with blue and red spheres, respectively.

The all- $\alpha \rightarrow$ all- β transition of RfaH-CTD is interesting for several reasons. Firstly, the genes in RfaH-regulated operons encode several bacterial virulence factors, including lipopolysaccharide (LPS) core, exopolysaccharide and haemolysin toxin, and the action of RfaH increases the expression of these factors, which are otherwise poorly transcribed (due to the large percentage of rare codons) and translated (due to a lack of canonical riboso-

mal recruitment sites). Hence, RfaH-CTD represents a good model for understanding gene regulation of these operons, which may be shared by other regulation factors. Secondly, the rules governing the refolding of RfaH-CTD may also be implicated in protein misfolding, so elucidating the mechanism for the large-scale structural transition of RfaH-CTD may aid in protein engineering and drug design.

While several experimental studies have been successful in characterising the domain-closed and domain-opened states of RfaH, the details of the refolding process have been inherently difficult to probe. NMR studies of the full-length RfaH are complicated by precipitation of the hydrophobic NTD once the protein dissociates from RNAP, or once domain dissociation is initiated *in vitro*.¹⁴ Additionally, NMR shifts for the isolated CTD strictly mirror those of the β -barrel conformer,¹⁴ and conversion back to the α -helical structure is not observed. This effect is largely due to the fact that contacts with the NTD are critical for inducing refolding to the α -helical state.¹⁶ Therefore, several groups have implemented computer simulation techniques to analyse the refolding process.

Unfortunately, large-scale structural rearrangements generally occur on relatively long time scales, and so are difficult to simulate in an efficient manner via standard techniques. The refolding of RfaH-CTD has been probed using several computational approaches, including replica exchange molecular dynamics (REMD),¹⁷ construction of Markov state models (MSMs),¹⁸ and replica-exchange-with-tunnelling (RET).¹⁹

REMD has been used to investigate the refolding of the isolated RfaH-CTD in implicit solvent.¹⁷ A free energy surface was constructed by projecting the replicas simulated at 310 K onto the root-mean-square deviation (rmsd) from the all- α state and the end-to-end distance. The structural transition was reported to proceed via a completely unfolded state, and the simulation yielded a relatively flat all- β -sheet structure compared to the barrel-like scaffold obtained in the NMR experiments.¹⁴ Li *et al.*¹⁸ constructed an MSM for RfaH-CTD from numerous MD trajectories. Based on the final MSM, they concluded that the conversion process could occur via heterogeneous routes, and postulated that the underlying energy

landscape for refolding was ‘rough’, which we interpret in terms of competing low energy structures separated by high barriers.

Recently, Bernhardt and Hansmann applied RET to decipher the refolding mechanism for RfaH-CTD.¹⁹ In RET,^{20,21} replicas evolve in the microcanonical ensemble for a short period, and are then provisionally exchanged, while simultaneously rescaling their velocities to ensure that the total energy is invariant. The replicas are then allowed to evolve again at constant energy, and the final structures are accepted or rejected based on a Metropolis criterion. This procedure ultimately may lead to improved acceptance probabilities compared to the standard REMD procedure. Using RET, a significant free energy barrier (approximately 10 RT) separating the all- α and all- β states of RfaH-CTD was identified, and the transition was reported to occur via a disordered conformer.¹⁹

In the present work, the potential energy landscape (PEL) framework and kinetic transition network (KTN) analysis are combined to probe the refolding of RfaH-CTD. In particular, discrete path sampling (DPS)^{22–24} is used to construct the PEL (which encompasses low-lying minima and the corresponding transition states that connect them) for the structural transition at full atomistic resolution. The free energy landscape (FEL) for RfaH-CTD is then derived from the PEL avoiding low-dimensional projections, and mechanistic details of the refolding process are outlined. We find that the free energy landscape of isolated RfaH-CTD at 310 K is multifunnelled. Consistent with previous NMR studies, the β -barrel state is more stable than the α -helical hairpin ensemble. We find that the structural transition occurs via a compact coil-like intermediate, and complete loss α -helical character.

2 Methods

Preparation of initial structures

The crystal structure of RfaH (residues 1–100 and 115–156) and the NMR solution structure of the isolated C-terminal region of RfaH (residues 97–162) were obtained from the protein

data bank via the PDB accession codes 2OUG¹³ and 2LCL,¹⁴ respectively. Residues 115–162 of the NMR structure were selected as the initial all- β conformer. The terminal six residues (157–162) were added to residues 115–156 of the crystal structure using PyMOL,²⁵ and the resulting structure represented the initial all- α conformer for the MD simulations.

Explicit solvent MD

The atomic interactions were modelled using the AMBER ff99SB-ILDN²⁶ parameter set. The initial all- α and all- β structures were first minimised in vacuum for 10000 steps. Each structure was then solvated using TIP3P water²⁷ in a truncated octahedron, with the box edges restricted to a minimum distance 10 Å from the protein. The solvated systems were then minimised for a further 10000 steps, and a restraining force of 100 kcal mol⁻¹ Å⁻² was applied to each protein structure. They were then heated from 0 to 300 K over 20 ps, with a weak restraint of 10 kcal mol⁻¹ Å⁻² on the protein. The restraints were subsequently removed and each system was equilibrated in the NPT ensemble (pressure = 1 atm; temperature = 300 K) for 5 ns, followed by 2 ns of constant volume MD. Finally, a 300 ns production run was performed at 300 K in the canonical ensemble. For all MD runs the temperature was regulated using a Langevin thermostat with a collision frequency of 1 ps. All bonds involving hydrogen were constrained using SHAKE, permitting a time step of 2 fs. Structures were saved every 10 ps for further analysis.

Preliminary analysis of the MD trajectories revealed that the all- β conformer sampled the native basin throughout the simulation (with relatively small deviations from the NMR topology). The structure with the lowest energy was selected as the starting geometry for discrete path sampling.^{22–24} For the all- α conformer, significant structural fluctuations were observed on the simulation time scale. Hence, additional equilibration and production (50 ns) runs were conducted for the α -helical hairpin, using backbone dihedral angle restraints, based on the crystal structure (i.e. for residues 115–156). The lowest energy α -helical conformers from both sets of MD runs were chosen as endpoints for DPS.

Construction of the potential energy landscape with DPS

To improve the efficiency of discrete path sampling, a generalised Born implicit solvent, GB-Neck2,²⁸ was used, with a cutoff 25 Å for evaluating the Born radius and an infinite cutoff for estimating non-bonded interactions. A salt concentration of 0.1 M was maintained, and the ff99SB-ILDN force field was also properly symmetrised, using the method suggested by Małolepsza *et al.*²⁹

DPS²²⁻²⁴ was performed using the OPTIM³⁰ and PATHSAMPLE³¹ programs with a GPU interface for OPTIM to accelerate sampling.³² Firstly, we obtained paths connecting the two α -helical conformers to the β -barrel scaffold (suggested by the MD simulations). Once two endpoints were chosen, a structural alignment was performed, which minimises the distance between them based on overall rotation, translation and permutation of identical atoms. The next step involves interpolation between the aligned configurations.

Since the conformational transition from the α -helical hairpin to the β -barrel scaffold is expected to be complex, RfaH-CTD is a good test system for the enhanced quasi-continuous interpolation (QCI)³³ scheme. Here, an auxiliary potential is used to derive a set of discrete images between two endpoints. The auxiliary potential contains constraint and repulsive terms for bonded and non-bonded atoms, respectively, and, in the latest scheme, sequence information from the AMBER topology file is employed. The new QCI routine also includes harmonic springs between images, and cis-trans peptide bond constraints. These improvements together minimise the likelihood of chain-crossings and cis-trans isomerism, which are undesirable consequences of (linear) interpolation techniques, especially for distant conformations.

The auxiliary potential is set up for the aligned endpoints, and discrete images between these two starting configurations are built by adding one atom (or residue) at a time. Before another atom (or residue) is added, the potential is minimised using an L-BFGS minimiser^{34,35} with a predefined root-mean-square (rms) gradient condition. This procedure is repeated until the full set of intermediate configurations is obtained for all atoms. The

minimised images were then used to seed a double-nudged³⁶ elastic band^{37,38} (DNEB) computation, which yields transition state guesses that are then tightly converged using hybrid eigenvector-following (HEF).^{39,40}

For a given set of RfaH-CTD endpoints, one QCI cycle was performed in the first connection attempt. DNEB-HEF cycles were then used for subsequent connection attempts. After each cycle, pairs of minima for connection were selected using a modified Dijkstra algorithm.⁴¹ To locate an initial path, this process was performed in parallel using the PATHSAMPLE program. The number of minima pairs to connect per cycle was defined *a priori*. The ‘best’ path between the two main endpoints was then computed using the Dijkstra missing connection algorithm.⁴¹ Unconnected minima on the best path were then chosen for QCI-DNEB-HEF/DNEB-HEF computations (QCI was only used if the minimised aligned distance between a given minima pair exceeded 50 Å). Once the connection runs for minima pairs were completed, the new minima and transition states were added to the existing database of stationary points. Connection cycles were repeated until a connected path was found.

The initial database was then refined using the SHORTCUT (decreases the number of steps on the fastest path by reconnecting minima separated by a given number of transition states)⁴¹⁻⁴³ and UNTRAP (removes artificial kinetic traps by reconnecting minima of comparable energies that are separated by high barriers)⁴² procedures in PATHSAMPLE, and progress was monitored by checking for convergence of the $\alpha \rightarrow \beta$ rate constant and the heat capacity curve.

Estimating free energy and kinetics

The free energy landscape for RfaH was computed at 300 and 310 K using the harmonic superposition approximation (HSA).⁴⁴ A recursive regrouping procedure⁴⁵ was employed to cluster minima and transition states in the kinetic transition network (KTN) into free energy groups, based on a free energy threshold. The structural rearrangement pathways (‘fastest

paths’) were computed for the regrouped KTN using Dijkstra’s shortest path algorithm with suitable edge weights,⁴¹ and rate constants were computed using a graph transformation approach.⁴⁶

Computation of structural order parameters

Secondary structure analysis was performed using the DSSP algorithm.⁴⁷ The mass-weighted geometric root-mean-square deviation (rmsd) from selected minima/free energy groups, and the radius of gyration (R_g) of free energy groups were computed using the CPPTRAJ program in the AMBER tools package.⁴⁸ The CPPTRAJ software was also used to compute the total number of hydrogen-bonds in the various RfaH-CTD conformational states, with hydrogen-bond distance and angle cutoffs of 3.5 Å and 150°, respectively.

Finally, the computed potential and free energy landscapes were visualised using disconnectivity graphs.^{49,50}

3 Results and Discussion

3.1 MD simulations for the α -helical and β -sheet conformers

In the domain-closed X-ray crystal structure the C-terminal domain of RfaH assumes an α -helical hairpin conformation with two antiparallel α -helices, and an intervening turn region. When domain separation is triggered, the CTD is known to refold in a β -barrel scaffold, with five antiparallel β -strands. Molecular dynamics simulations (300 ns) were used to probe the short time stability of the two extreme RfaH-CTD forms.

The simulation initiated from the α -helical conformer (Figure 2a) shows that α_4 (residues 115–140; see Figure 1) has a higher propensity for helical unwinding than α_5 (residues 135–155); α_4 is partially unfolded throughout the entire production run, while α_5 maintains most of its α -helical structure. These findings agree well with previous work,^{17,18} in which α_4 was reported to be less stable than α_5 for the isolated CTD. Several authors^{16,51} suggest

that interdomain contacts between the NTD and the CTD are responsible for maintaining the stability of the α -helical form of RfaH-CTD, and when these contacts are disrupted the probability of forming the β -sheet analogue increases.

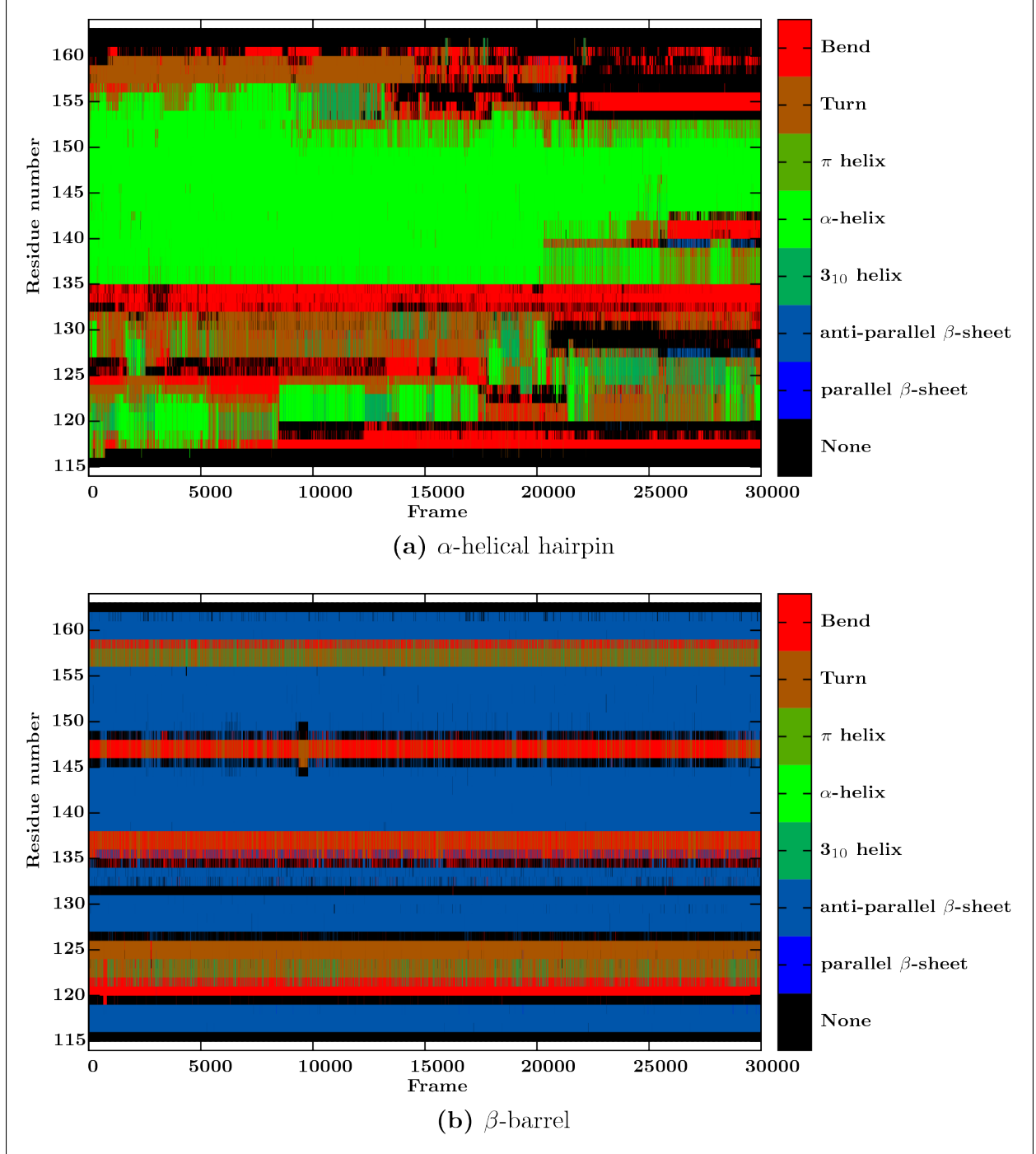


Figure 2: Secondary structure assignments for configurations along MD trajectories. The MD simulations (300 ns) for RfaH-CTD α -helical hairpin and β -barrel conformers were computed at 300 K in the NVT ensemble with explicit solvent.

Figure 2b reveals that the β -barrel form of isolated RfaH-CTD is quite stable on the short

MD simulation time scale. Throughout the MD run, the β -strands remain intact; with β_2 (127–130), β_3 (138–144) and β_4 (149–155) closely matching the NMR solution structure,¹⁴ and β_1 (116–119) and β_5 (159–161) starting one residue later. An additional short β -strand (residues 132–133) was predicted between β_2 and β_3 . A previous study also found that these residues had a tendency to adopt β -sheet structure, predicting that β_2 extended from residues 127 to 134.¹⁷

The MD β -barrel and the partially unfolded α -helical structure are likely to be important conformers on the potential energy landscape for RfaH-CTD. They were therefore chosen as endpoints for discrete path sampling. However, since we are mainly interested in probing the refolding process from the α -helical hairpin form, we performed further structural refinement of the crystal structure with dihedral angle restraints (Figure 2). The refined structure was also used as an endpoint for DPS. These initial DPS endpoints closely resemble the structures depicted in Figure 5a (see section 3.3).

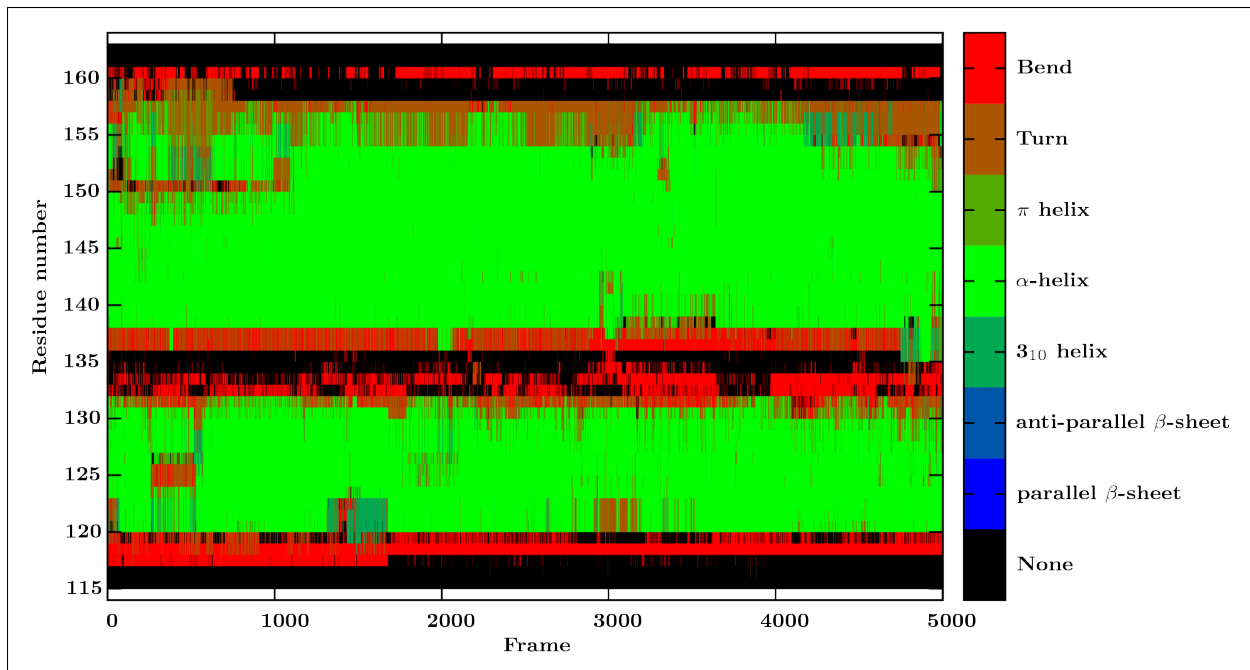


Figure 3: Secondary structure assignments for configurations along MD trajectory. The MD simulation (50 ns) for RfaH-CTD α -helical hairpin was computed at 300 K in the NVT ensemble with explicit solvent. Backbone dihedral angle restraints for residues 115–156 were employed throughout.

3.2 Effects of QCI parameters on optimised paths

Computation of initial pathways between selected endpoints represents one of the main challenges in DPS. An interpolation procedure is first used to predict intervening structures between a given pair of conformers, which are then optimised to yield transition states and corresponding local minima, as described in section 2. For conformers close in configuration space, an initial linear interpolation scheme is generally sufficient; however, such schemes perform poorly for distant minima.

The quasi-continuous interpolation (QCI) scheme^{33,52} has recently been shown to yield kinetically relevant paths for several large-scale rearrangements.^{33,53} It allows the user to control several parameters; including the total number of images (i.e. intervening geometries; N_{\max}^{images}), the cutoff distance for activating repulsive terms between unconstrained atoms

(r_{rep}), the force constant for harmonic springs connecting images (k_{spr}), and the method used for growing the images (e.g. atom-by-atom, residue-by-residue), among others.

Table 1: Comparison of selected QCI parameters for two different interpolations.

QCI parameters	Int-I	Int-II
r_{rep} (Å)	8.0	6.0
$k_{\text{spr}}^{\text{images}}$	10.0	10.0
$N_{\text{max}}^{\text{images}}$	200	50
method	add residue	add residue

Table 1 compares some QCI parameters for two different interpolations from the RfaH-CTD lowest MD α conformer to the lowest MD β structure. In the first interpolation (Int-I) more images were used ($N_{\text{max}}^{\text{images}} = 200$) and a slightly larger repulsive cutoff distance was employed ($r_{\text{rep}} = 8.0$) than in Int-II. In both schemes the same value was set for the spring force constant, and images were constructed by adding one residue at a time.

The resulting optimised initial path corresponding to each QCI interpolation scheme is depicted in Figure 4. Int-II leads to a significantly shorter path connecting the α and β conformers than the final path obtained using Int-I. In the latter case, the protein becomes kinetically trapped, over about 1000 steps, before finally folding downhill towards the β -sheet structure. In this case, it seems that having a large number of images is actually less efficient. Interestingly, when the two paths were merged into one KTN, the longer path was no longer kinetically competitive. Hence, it would be beneficial to include initial paths corresponding to different QCI interpolations (for the same set of endpoints) in the KTN, to increase the likelihood of finding the most biologically relevant paths.

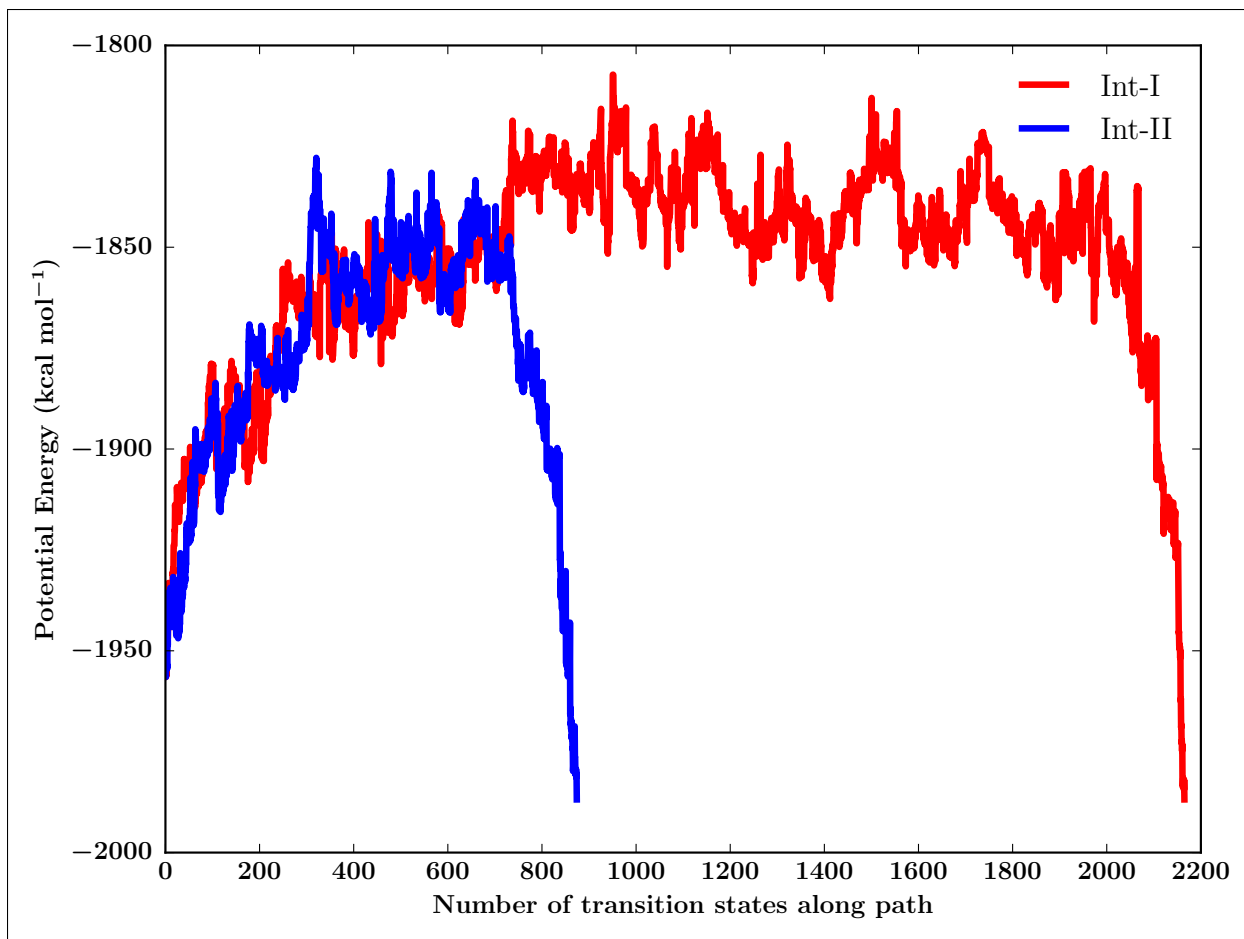


Figure 4: Optimised initial paths corresponding to two different QCI interpolations from the lowest- α to lowest- β conformer.

3.3 Potential and free energy landscapes

The PEL for the isolated RfaH-CTD is shown in Figure 5a; there are two prominent deep funnels. The major funnel, which includes the all- β conformer, is notably lower in energy than the one corresponding to the all- α structure, and contains the global minimum. The partially unfolded α -helical conformer is enthalpically more favourable than the α -helical hairpin form.

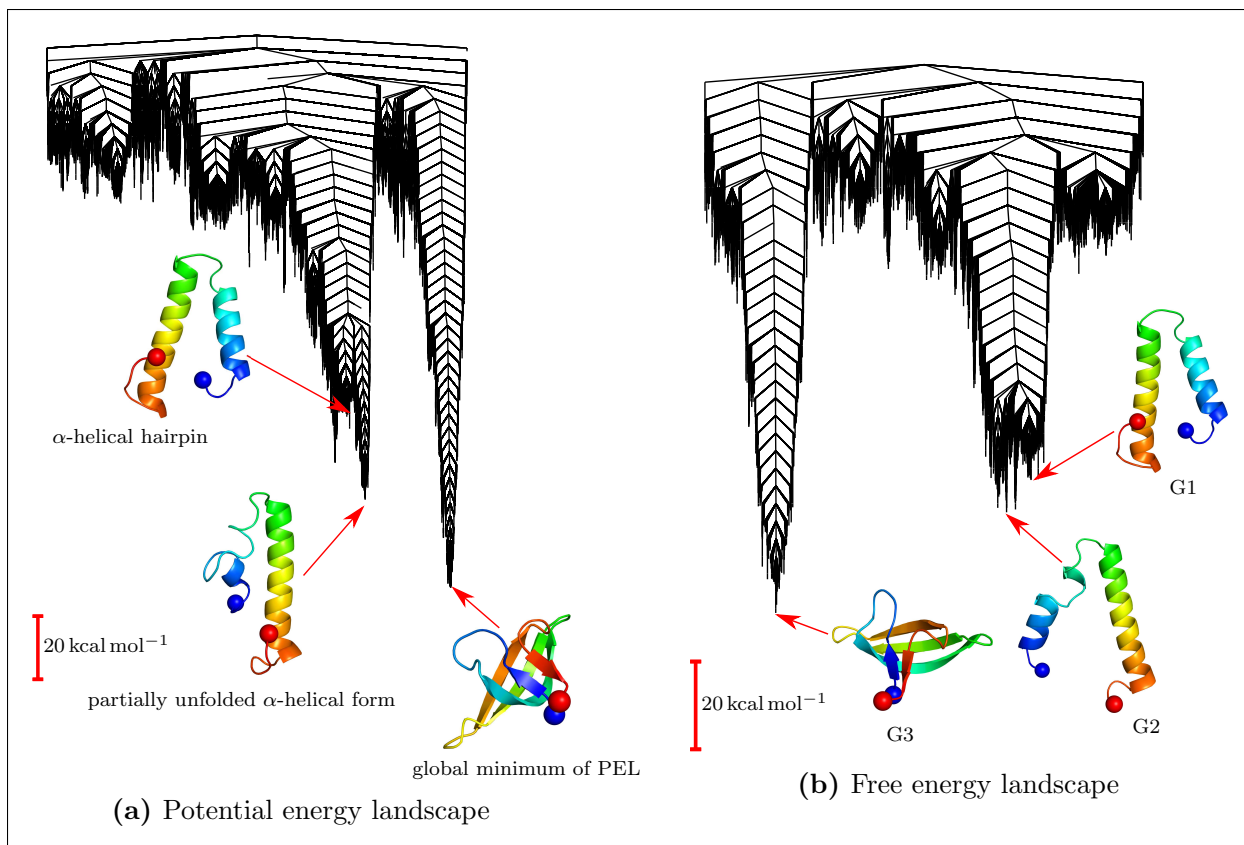


Figure 5: Disconnectivity graphs for the isolated RfaH-CTD, in terms of (a) potential and (b) free energies. In (a) the lowest energy α -helical conformer (partially unfolded), the α -helical conformer with maximum helical content (hairpin), and the global minimum of the PEL are all superimposed on the graph. The free energies were computed at 310 K with minima and transition states regrouped⁴⁵ based on an energy threshold of 5 kcal mol⁻¹. Representative structures for selected free energy groups (G1 to G3) are also shown.

The FEL was computed from the PEL at 300 K (not shown) and 310 K (Figure 5b). These two temperatures were chosen to allow for direct comparison with previous simulations (MSM construction at 300 K;¹⁸ replica exchange approaches at 310 K^{17,19}) and the original NMR experiment (at 310 K).¹⁴ There was no significant difference between the two the landscapes, and so further analysis refers to the FEL at 310 K. Each branch on the free energy disconnectivity graph corresponds to a free energy group. The topology of the global free energy minimum, G3, is consistent with the NMR solution structure for the isolated RfaH-CTD (all-atom geometric rmsd = 1.57 Å); however, β_2 is visibly longer compared to

the experimental structure. From the FEL, it is evident that the β -barrel scaffold is the most stable conformer for the isolated CTD of RfaH. In addition, the partially unfolded α -helical state, G2, is slightly more stable than the analogue with both helices intact, G1. Combined with the MD results, these results suggest that upon domain separation α_4 quickly loses some of its helical character, and G2 is an important intermediate in the refolding process of RfaH-CTD.

Since the barriers on the FEL for the isolated domain are particularly high, we infer that, in the absence of the appropriate molecular partner, the refolding process is likely to be slow. In fact, Burmann *et al.* probed the refolding process, by engineering a cleavage site into the linker region between the two domains, and reported that β -sheet structure was only detected 42 hours after incubation.¹⁴

3.4 Conformational states on the FEL

To gain better insight into the various conformational states on the FEL, the free energy disconnectivity graph was coloured based on several structural order parameters. Secondary structure analysis was performed for each free energy group, and these results are summarised in Figures 6a–c. Considerable variation in α -helical and β -sheet content is observed in Figures 6a and b. The G1 ensemble displays about 77% α -helical content, while ensembles in the high energy regions of the FEL and in the neighbourhood of the global free energy minimum (G3) generally show negligible α -helical character. Maximum β -sheet content was observed for G3 (68%), and ensembles in the intermediate regions of the FEL contain some degree of α -helical or β -sheet content. Significant coil-like structure (i.e. lack of regular secondary structure) was observed for many ensembles in the high energy region of the landscape (e.g. G11 in Figure 6c).

The free energy disconnectivity graphs are also depicted in terms of the all-atom geometric rmsd from G1 (Figure 6d) and G3 (Figure 6e). These graphs further highlight the inherent structural heterogeneity of the states on the FEL. The principal funnel correspond-

ing to ensembles with high α -helical content separates into two main sub-funnels: ensembles closely related to the hairpin state (G1) and those with α_4 partially unfolded (e.g. G7). The ensembles gradually deviate from G1 on traversing the landscape towards G3. A similar trend is observed from G3 towards G1.

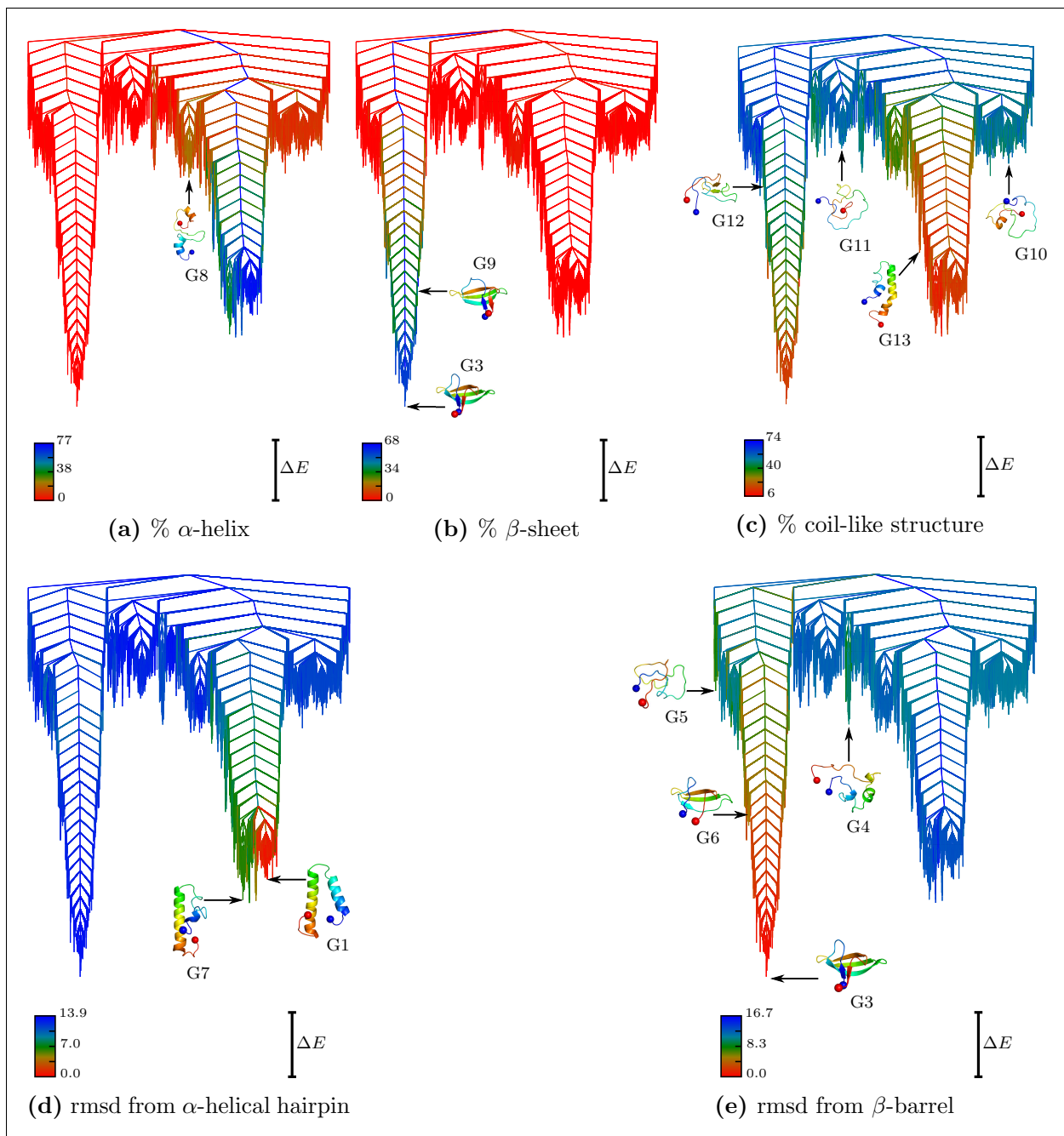


Figure 6: Free energy disconnectivity graphs ($\Delta E = 20 \text{ kcal mol}^{-1}$) for RfaH-CTD computed at 310 K with a regrouping⁴⁵ threshold of 5 kcal mol^{-1} . The landscape is reproduced for several structural order parameters, and representative structures for selected free energy groups are highlighted.

Based on these results, we infer that on moving from the α -helical hairpin ensemble to the β -barrel state, RfaH-CTD gradually loses α -helical character, and the structural conversion

occurs via an essentially unstructured intermediate.

3.5 Mechanism for fold-switching in RfaH-CTD

A detailed description of the refolding process can be obtained by examining the pathway between the all- α and all- β conformations that makes the largest contribution to the rate constant. For RfaH-CTD, the two forms were again defined as states by lumping stationary points into free energy groups, using recursive regrouping⁴⁵ with an energy threshold of 11 kcal mol⁻¹. Regrouping thresholds for which the rate constant is converged give consistent results. However, if the threshold is too small excessive detail may be retained, and analysing the mechanism may prove difficult. The fastest pathway between selected states was extracted by employing Dijkstra’s shortest path algorithm on the clustered stationary point database.⁴¹

Figure 7 shows the variation in several structural order parameters along the fastest pathway. A significant deviation from the initial α -helical hairpin coincides with helical unwinding of α_4 (Figure 7a); GLU124 to THR131 unfolds and a short turn develops (GLN127 to ALA128). Two other groups also reported that unwinding of α_4 marked the first stage of the structural transition.^{17,18} Geometric rms deviations from the all- α state oscillate at around 7 Å for about eight steps; α_4 continues to shorten, while α_5 generally remains intact. The protein then passes through an ‘unstructured’ intermediate (at step 19; Figure 7a), and then the configurations progressively become more β -like. Li *et al.* also observed a high population of compact coil-like states in their MSM for RfaH-CTD.¹⁸

The α -helical content decreases sharply at step ten of the folding transition (Figure 7b). At this stage, a transition state develops with low helical content in α_5 , only maintaining helical structure from ALA137 to LEU142. From steps 14 to 19 the protein contains negligible α -helix or β -sheet structure; in that part of path states display maximum coil-like structure (notice the green curve in Figure 7b). The protein only adopts β -sheet-like structure in the latter segment of the path, as the canonical β -strands begin to form.

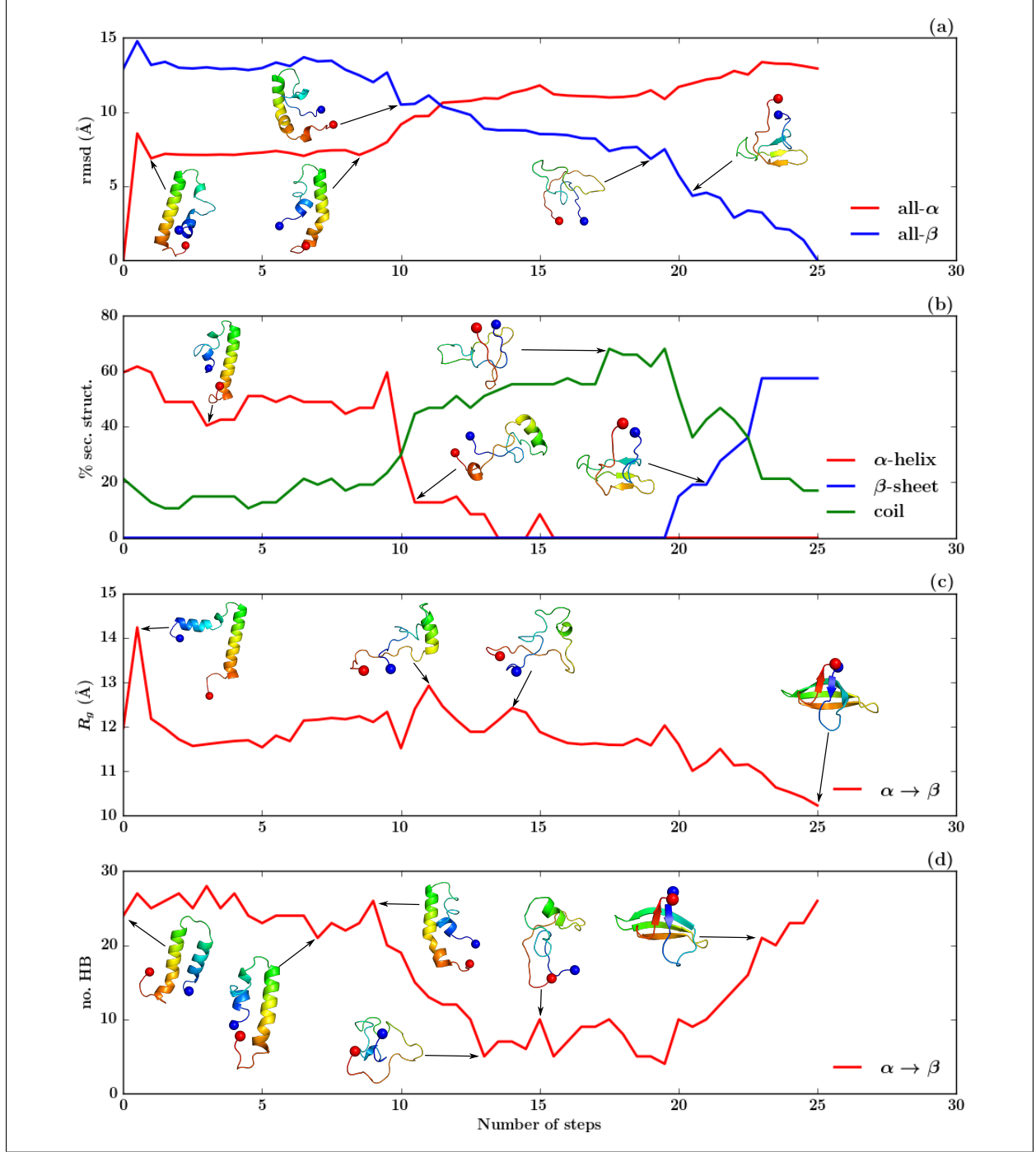


Figure 7: Evolution of selected order parameters on the fastest folding pathway from the α -helical hairpin to β -barrel state of RfaH-CTD: (a) all-atom geometric rmsd from all- α or all- β state; (b) secondary structure content; (c) radius of gyration; (d) total number of hydrogen-bonds. Representative structures for some states along the path are shown. The number of steps corresponds to the number of transition states along the path.

The radius of gyration (R_g), which was taken as the average mass-weighted squared distances of all atoms from protein centre of mass, is another useful order parameter for monitoring structural changes during the rearrangement process. For most of the refolding process R_g is about 12 Å, suggesting that the protein remains relatively compact during the transition (Figure 7c). Notably, in the early stages of folding, a significant increase in R_g is observed; a transition state forms with the two α -helices orientated roughly orthogonal to each other. This state is strikingly similar to the one located on the free energy surface of RfaH-CTD by GC *et al.*¹⁷ On further investigation, it seems that this state forms due to the breakage of a hydrogen-bond between THR119:HG1 (in α_4) and GLU149:O (in α_5), which causes the two helices to separate. However, new hydrogen-bonds are formed; for example, a short turn simultaneously forms in α_4 , perhaps to accommodate the nearby bulky phenylalanine residues (PHE126, PHE130).

The variation in the number of hydrogen-bonds along the path was also examined. Figure 7d reveals that the protein does maintain some degree of hydrogen-bonding throughout refolding. For instance, between steps 13 to 19 (when the coil-like character is at a maximum) there is still some hydrogen-bonding due to turns (e.g. PRO133 to ASP134) and 3_{10} helices (e.g. GLY125 to GLN127). The intermediates in that region also contain a significant number of bends (loops), which lead to compact morphologies. The hydrogen-bond pattern then increases steadily as the β -strands nucleate to yield the all- β state.

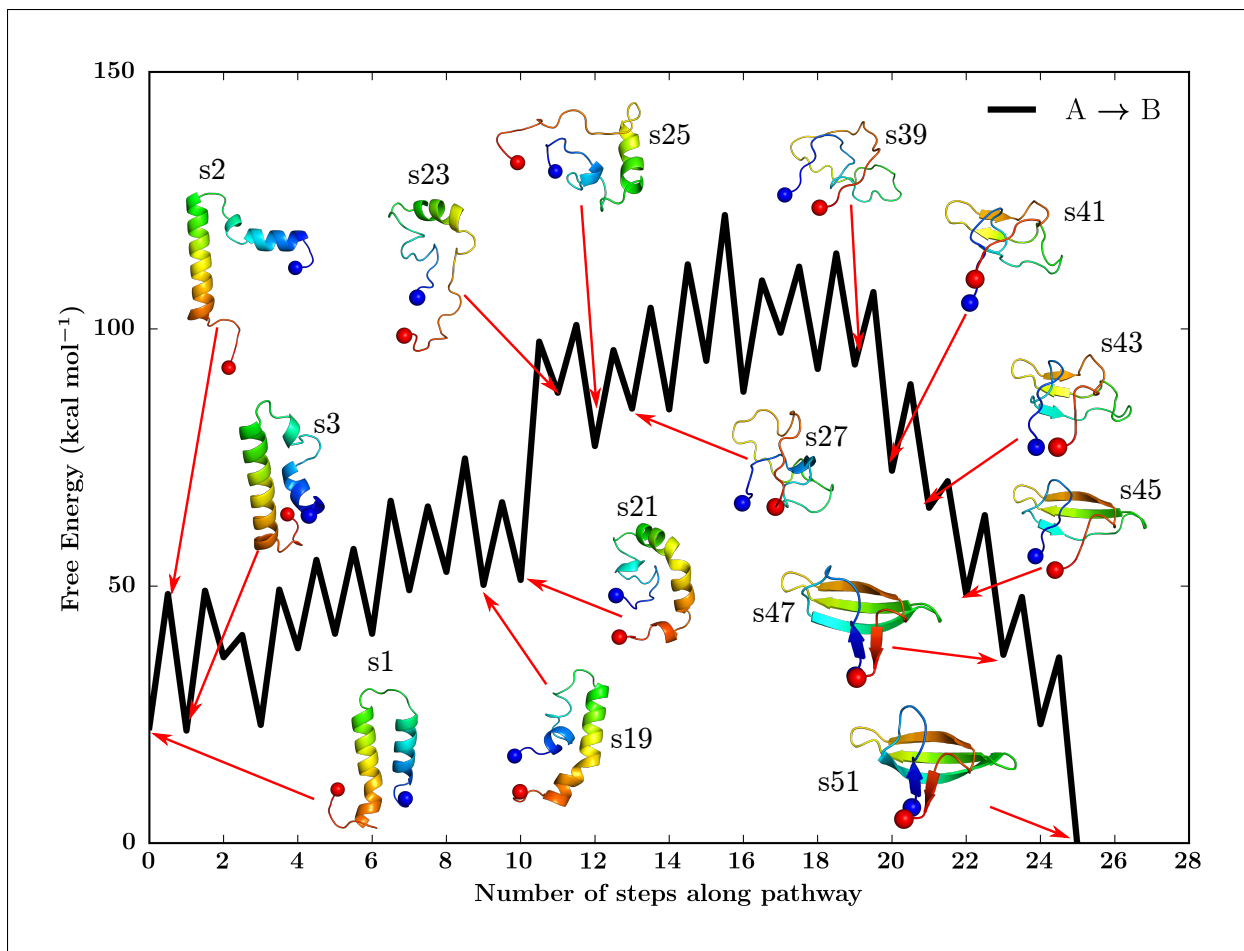


Figure 8: Pathway for the α -helical hairpin \rightarrow β -barrel structural rearrangement of RfaH-CTD. Stationary points in the kinetic transition network were regrouped based on a threshold of 11 kcal mol^{-1} . Representative structures of selected states are superimposed on the path. States are numbered based on their positions along the path: s1 corresponds to the α -helical hairpin ensemble, and s51 represents the β -barrel state.

Finally, the refolding pathway of RfaH-CTD is presented in Figure 8 in terms of free energies. The structural transition occurs in three main stages:

1. The formation of a kink (short turn) in the neighbourhood of the bulky phenylalanine residues initiates the refolding process (s3). α_4 gradually shortens in the direction of the N-terminus (s19). Loss of α -helical character in α_4 then accommodates expansion of α_5 (s21)—starting from the C-terminus.

2. Helical unwinding eventually leads to the formation of a compact intermediate (s23), which includes residual α -helical character (ALA137 to LEU142). The formation of this intermediate is preceded by a major free energy barrier in the refolding process, which may therefore be classified as rate-limiting. Once unfolding of α_5 is complete, the C-terminal part of the protein crosses over the N-terminus, yielding a compact coil-like state (s27). Small conformational changes lead to the formation of s39, which exhibits a β -barrel-like topology, with complete loss of α -helical character.
3. Once s39 forms, nucleation of the β -strands commences. β_3 (LEU142 to ASN144) and β_4 (GLU149 to LYS151) begin to develop first (s41), followed by nucleation of β_2 (s43). Strands 1 and 5 form last—completing the β -scaffold (s51).

4 Conclusions

Large-scale conformational changes in proteins are relatively difficult to probe. Such structural transformations may lead to the exposure of hydrophobic residues, resulting in aggregation *in vitro*, impeding experimental characterisation. Additionally, one metamorphic partner may be more stable than the other, and so probing the reverse process, at physically relevant temperatures, may be a challenge.

Computer simulations of fold-switching can therefore play an important role in improving our understanding of these processes, and aid in the design of novel protein-based architectures. However, *in silico* studies of systems undergoing large-scale changes have their own challenges. In particular, these processes often occur on long time scales, and the morphologies of interest may be separated by substantial free energy barriers. To circumvent these issues various sampling and data analysis strategies have been adopted.

In the present work, methods based on geometry optimisation were employed to characterise energy landscapes for the all- α to all- β transition of RfaH-CTD. The new quasi-continuous interpolation scheme³³ was employed to obtain initial guesses for putative struc-

tures on the refolding path; together with other discrete path sampling strategies, a kinetic transition network for RfaH-CTD was constructed consisting of stationary points on the potential energy landscape.

The free energy landscape for RfaH-CTD was computed at 310 K within the harmonic superposition approximation. The landscape is characteristically multifunnelled, and, consistent with experiment,¹⁴ the β -barrel scaffold is the favoured conformer. The proposed mechanism for the structural transition is in good agreement with previous work,¹⁸ and some of the important structural ensembles identified in this study have been found in REMD simulations¹⁷ and in MSM constructions.¹⁸ New atomistic details for the refolding process have been provided in the present work.

The ability of our approach to preserve the full atomistic resolution should aid in deriving design principles for protein fold-switching. It would therefore be of interest to extend this work to other transformer proteins (as they become available) and related systems.

Acknowledgement

J. A. J acknowledges financial support from the Gates Cambridge Trust. D. J. W gratefully acknowledges support from the EPSRC.

References

- (1) Littler, D. R.; Harrop, S. J.; Fairlie, W. D.; Brown, L. J.; Pankhurst, G. J.; Pankhurst, S.; DeMaere, M. Z.; Campbell, T. J.; Bauskin, A. R.; Tonini, R. The intracellular chloride ion channel protein CLIC1 undergoes a redox-controlled structural transition. *J. Biol. Chem.* **2004**, *279*, 9298–9305.
- (2) Luo, X.; Tang, Z.; Xia, G.; Wassmann, K.; Matsumoto, T.; Rizo, J.; Yu, H. The Mad2

- spindle checkpoint protein has two distinct natively folded states. *Nat. Struct. Mol. Biol.* **2004**, *11*, 338.
- (3) Andreeva, A.; Murzin, A. G. Evolution of protein fold in the presence of functional constraints. *Curr. Opin. Struct. Biol.* **2006**, *16*, 399–408.
 - (4) Meier, S.; Özbek, S. A biological cosmos of parallel universes: does protein structural plasticity facilitate evolution? *BioEssays* **2007**, *29*, 1095–1104.
 - (5) Murzin, A. G. Metamorphic proteins. *Science* **2008**, *320*, 1725–1726.
 - (6) Tuinstra, R. L.; Peterson, F. C.; Kutlesa, S.; Elgin, E. S.; Kron, M. A.; Volkman, B. F. Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5057–5062.
 - (7) Tokuriki, N.; Tawfik, D. S. Protein dynamism and evolvability. *Science* **2009**, *324*, 203–207.
 - (8) Bryan, P. N.; Orban, J. Proteins that switch folds. *Curr. Opin. Struct. Biol.* **2010**, *20*, 482–488.
 - (9) Bailey, M. J. A.; Koronakis, V.; Schmoll, T.; Hughes, C. Escherichia coli HIyT protein, a transcriptional activator of haemolysin synthesis and secretion, is encoded by the rfaH (sfrB) locus required for expression of sex factor and lipopolysaccharide genes. *Mol. Microbiol.* **1992**, *6*, 1003–1012.
 - (10) Bailey, M. J. A.; Hughes, C.; Koronakis, V. RfaH and the ops element, components of a novel system controlling bacterial transcription elongation. *Mol. Microbiol.* **1997**, *26*, 845–851.
 - (11) Bailey, M. J. A.; Hughes, C.; Koronakis, V. In vitro recruitment of the RfaH regulatory protein into a specialised transcription complex, directed by the nucleic acid ops element. *Mol. Gen. Genet. MGG* **2000**, *262*, 1052–1059.

- (12) Belogurov, G. A.; Mooney, R. A.; Svetlov, V.; Landick, R.; Artsimovitch, I. Functional specialization of transcription elongation factors. *EMBO J.* **2009**, *28*, 112–122.
- (13) Belogurov, G. A.; Vassylyeva, M. N.; Svetlov, V.; Klyuyev, S.; Grishin, N. V.; Vassylyev, D. G.; Artsimovitch, I. Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol. Cell* **2007**, *26*, 117–129.
- (14) Burmann, B. M.; Knauer, S. H.; Sevostyanova, A.; Schweimer, K.; Mooney, R. A.; Landick, R.; Artsimovitch, I.; Rösch, P. An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* **2012**, *150*, 291–303.
- (15) Knauer, S. H.; Artsimovitch, I.; Rösch, P. Transformer proteins. *Cell Cycle* **2012**, *11*, 4289–4290.
- (16) Tomar, S. K.; Knauer, S. H.; NandyMazumdar, M.; Rösch, P.; Artsimovitch, I. Inter-domain contacts control folding of transcription factor RfaH. *Nucleic Acids Res.* **2013**, *41*, 10077–10085.
- (17) GC, J. B.; Bhandari, Y. R.; Gerstman, B. S.; Chapagain, P. P. Molecular dynamics investigations of the α -helix to β -barrel conformational transformation in the RfaH transcription factor. *J. Phys. Chem. B* **2014**, *118*, 5101–5108.
- (18) Li, S.; Xiong, B.; Xu, Y.; Lu, T.; Luo, X.; Luo, C.; Shen, J.; Chen, K.; Zheng, M.; Jiang, H. Mechanism of the all- α to all- β conformational transition of RfaH-CTD: Molecular dynamics simulation and markov state model. *J. Chem. Theory Comput.* **2014**, *10*, 2255–2264.
- (19) Bernhardt, N. A.; Hansmann, U. H. E. Multi-funnel landscape of the fold-switching protein RfaH-CTD. *J. Phys. Chem. B* **2018**,
- (20) Yaar, F.; Bernhardt, N. A.; Hansmann, U. H. E. Replica-exchange-with-tunneling for fast exploration of protein landscapes. *J. Chem. Phys.* **2015**, *143*, 224102.

- (21) Bernhardt, N. A.; Xi, W.; Wang, W.; Hansmann, U. H. E. Simulating protein fold switching by replica exchange with tunneling. *J. Chem. Theory Comput.* **2016**, *12*, 5656–5666.
- (22) Wales, D. J. Discrete path sampling. *Mol. Phys.* **2002**, *100*, 3285–3305.
- (23) Wales, D. J. Some further applications of discrete path sampling to cluster isomerization. *Mol. Phys.* **2004**, *102*, 891–908.
- (24) Evans, D. A.; Wales, D. J. Folding of the GB1 hairpin peptide from discrete path sampling. *J. Chem. Phys.* **2004**, *121*, 1080–1090.
- (25) The PyMOL molecular graphics system, version 1.8. 2015.
- (26) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E.; Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.
- (27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (28) Nguyen, H.; Roe, D. R.; Simmerling, C. Improved generalized born solvent model parameters for protein simulations. *J. Chem. Theory Comput.* **2013**, *9*, 2020–2034.
- (29) Małolepsza, E.; Strodel, B.; Khalili, M.; Trygubenko, S.; Fejer, S. N.; Wales, D. J. Symmetrization of the AMBER and CHARMM force fields. *J. Comput. Chem.* **2010**, *31*, 1402–1409.
- (30) Wales, D. J. OPTIM: A program for optimising geometries and calculating pathways. <http://www-wales.ch.cam.ac.uk/OPTIM>.

- (31) Wales, D. J. PATHSAMPLE: A driver for OPTIM to create stationary point databases using discrete path sampling and perform kinetic analysis. <http://www-wales.ch.cam.ac.uk/PATHSAMPLE>.
- (32) Mantell, R. G.; Pitt, C. E.; Wales, D. J. GPU-accelerated exploration of biomolecular energy landscapes. *J. Chem. Theory Comput.* **2016**, *12*, 6182–6191.
- (33) Röder, K.; Wales, D. J.; Roder, K.; Wales, D. J. Predicting pathways between distant configurations for biomolecules. *J. Chem. Theory Comput.* **2018**, *14*, 4271–4278.
- (34) Nocedal, J. Updating quasi-Newton matrices with limited storage. *Math. Comput.* **1980**, *35*, 773–782.
- (35) Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, *45*, 503–528.
- (36) Trygubenko, S. A.; Wales, D. J. A doubly nudged elastic band method for finding transition states. *J. Chem. Phys.* **2004**, *120*, 2082–2094.
- (37) Henkelman, G.; Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **1999**, *111*, 7010–7022.
- (38) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.
- (39) Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **1999**, *59*, 3969.
- (40) Kumeda, Y.; Wales, D. J.; Munro, L. J. Transition states and rearrangement mechanisms from hybrid eigenvector-following and density functional theory.: application

- to C10H10 and defect migration in crystalline silicon. *Chem. Phys. Lett.* **2001**, *341*, 185–194.
- (41) Carr, J. M.; Trygubenko, S. A.; Wales, D. J. Finding pathways between distant local minima. *J. Chem. Phys.* **2005**, *122*, 234903.
- (42) Strodel, B.; Whittleston, C. S.; Wales, D. J. Thermodynamics and kinetics of aggregation for the GNNQQNY peptide. *J. Am. Chem. Soc.* **2007**, *129*, 16005–16014.
- (43) Wales, D. J.; Carr, J. M.; Khalili, M.; de Souza, V. K.; Strodel, B.; Whittleston, C. S. *Proteins Energy, Heat Signal Flow*; Computation in chemistry; CRC Press, 2009; p 315.
- (44) Strodel, B.; Wales, D. J. Free energy surfaces from an extended harmonic superposition approach and kinetics for alanine dipeptide. *Chem. Phys. Lett.* **2008**, *466*, 105–115.
- (45) Carr, J. M.; Wales, D. J. Folding pathways and rates for the three-stranded β -sheet peptide Beta3s using discrete path sampling. *J. Phys. Chem. B* **2008**, *112*, 8760–8769.
- (46) Wales, D. J. Calculating rate constants and committor probabilities for transition networks by graph transformation. *J. Chem. Phys.* **2009**, *130*, 204111.
- (47) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogenbonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
- (48) Roe, D. R.; Cheatham III, T. E. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.
- (49) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

- (50) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760.
- (51) Ramírez-Sarmiento, C. A.; Noel, J. K.; Valenzuela, S. L.; Artsimovitch, I. Interdomain contacts control native state switching of RfaH on a dual-funneled landscape. *PLoS Comput. Biol.* **2015**, *11*, e1004379.
- (52) Wales, D. J.; Carr, J. M. Quasi-continuous interpolation scheme for pathways between distant configurations. *J. Chem. Theory Comput.* **2012**, *8*, 5020–5034.
- (53) Chakraborty, D.; Wales, D. J. Probing helical transitions in a DNA duplex. *Phys. Chem. Chem. Phys.* **2017**, *19*, 878–892.

Graphical TOC Entry

